DOCUMENT RESUME

ED 429 104 TM 029 642

AUTHOR Balizet, Sha; Treder, Dave; Parshall, Cynthia G.
TITLE The Development of an Audio Computer-Based Classroom Test of ESL Listening Skills.
PUB DATE 1999-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150) -- Tests/Questionnaires (160)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Audio Equipment; Audiovisual Communications; *College Students; *Computer Assisted Testing; Educational Technology; *English (Second Language); Higher Education; *Listening Comprehension Tests; Sound Effects; Speech; *Test Construction
IDENTIFIERS Paper and Pencil Tests

ABSTRACT

There are very few examples of audio-based computerized tests, but for many disciplines, such as foreign language and music, there appear to be many benefits to this type of testing. The purpose of the present study was to develop and compare computer-delivered and audiocassette/paper-and-pencil versions of a listening test. The test was a measure of progress achievement of academic listening comprehension and vocabulary for high-intermediate level students of English as a Second Language (ESL) at a university-affiliated institute. The underlying assumption investigated was that the use of computer and audio technology for classroom progress tests would provide benefits of convenience and improved sound quality while providing measurement quality and validity that were at least comparable to the paper form of such tests. Results with 28 students indicate that the computerized test performed at least as well as the paper-and-pencil version, with generally comparable validity. Appendixes contain the paper-and-pencil test and a list of the student survey questions. (Contains 4 figures, 11 references, and 7 additional resources.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 429 104

The Development of an Audio Computer-Based Classroom Test of ESL Listening Skills

Sha Balizet
Dave Treder
Cynthia G. Parshall
University of South Florida

The number of computer exams has increased dramatically in the last decade, especially for large-scale placement tests. However, there are currently very few applications of *audio*-based computerized tests. This limited development is especially true in terms of classroom or progress tests. For a wide range of disciplines, such as foreign language and music, there appear to be significant benefits to this type of computerized testing.

The purpose of the present research was to develop and compare computer-delivered and audiocassette/paper-and-pencil versions of a listening test. The test was a measure of progress achievement of academic listening comprehension and vocabulary for high-intermediate level students of English as a Second Language (ESL) at a university-affiliated institute. Our underlying assumption, investigated in this study, was that the use of computer and audio technology for classroom progress tests would provide benefits of convenience and improved sound quality, while providing measurement quality and validity that were at least comparable to the paper form of such tests.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Cynthia Parshall
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the American Educational Research Association, *Computer Applications in Education SIG*, Montreal, Canada, April 19-23, 1999.

TM029642

BEST COPY AVAILABLE

The Development of an Audio Computer-Based Classroom Test of ESL Listening Skills

Second and foreign language testing has advanced in many important ways since Lado's (1961) publication, *Language Testing*, which is considered the seminal work that established language testing as a field of specialization. Now the field is served by special interest academic journals, an electronic mail discussion list, and a website. (For further information, see *Additional Resources* at the end of this paper.)

Language testing is an area of specialization where many other fields come into play. Bachman (1990) noted that the present state of development in language testing incorporates the progress made in several other areas and that language testing holds a close, reciprocal relationship with research into language teaching and language acquisition.

Types of Language Tests

Language tests, like the field of language testing , are best viewed in context. Hughes (1989) identified four types of language tests. A *proficiency test,* such as the *Test of English as a Foreign Language* (TOEFL), measures examinees' language ability, and is broadly based. A *diagnostic test* identifies learners' strengths and the areas where they need to improve. Hughes' third test category is *placement tests*, which are typically used by institutions to determine incoming students' level or class placement. Commercially available placement tests include the *Comprehensive English Language Test* (CELT).

Hughes also identifies a fourth type of test, the *achievement test*. Unlike the other three kinds of tests, the achievement test typically bears a close connection to a given course. Hughes distinguishes between *final* achievement tests which summatively measure how well students have met course objectives and *progress* achievement tests, which evaluate how successfully students are advancing towards course objectives.

Language tests can be viewed according to the purposes they serve, as above, or they can be characterized by the specific skill areas that they measure. It is generally agreed that language consists of four skill areas: the receptive skills of listening and reading, and the productive skills of speaking and writing. These skill areas are often taught in combinations such as L/S and R/W; or, in the integrated approach, all four skills are included.

Listening Tests

The test developer faces special challenges when creating tests of listening comprehension. A primary problem is that listeners do not demonstrate measurable, overt behavior. More fundamentally, assessment cannot be guided by a broad, overarching theory of listening behavior: no such theory exists (Powers, 1986). Whatever is known about listening results from research with native speakers. There is a need for greater conceptual clarity and more research on the listening of non-native speakers or second language (L2) learners (Richards, 1983).

Regardless of limitations in listening theory, there remains a need for assessing learners' L2 listening comprehension. Instructors of foreign and second languages often focus their progress tests on such skills as *listening for specific information*, *identifying main ideas*, and *recognizing the meaning of intonation patterns*. Another approach, appropriate for university bound L2 learners, would be to measure the *lecture-listening skills* that they will need in their academic futures.

Computer-Adaptive Language Tests

Until recently, language tests typically used a paper-and-pencil format, with listening prompts delivered by audiocassette. However, computers are increasingly more powerful and available; as a consequence, considerable work has been done on developing computerized language tests. Most of these have been measures of reading skills, or tests of grammar structures or vocabulary.

Substantial attention has been paid to computer-adaptive language tests (CALTs), where each examinee receives a test uniquely tailored to his/her ability level. The U.S. Department of Education funded development of the first CALT, issued in 1986 at Brigham Young University (Madsen, 1991). This placement test has reading and listening components. Another CALT is used for entry and exit decisions with limited-English proficient (LEP) grade school children in the Montgomery County (Maryland) Pubic Schools (Stevenson and Gross, 1991). Another major CALT development effort by Ariew and Dunkel (1989) was funded by the U.S. Department of Education and backed by the American Council on the Teaching of Foreign Languages (ACTFL). Ariew and Dunkel constructed prototype computerized tests of listening proficiency in French and in English as a second language. A final example of computerized adaptive language tests is the new TOEFL, which has supplanted the former paper-and-pencil version in the United States (Educational Testing Service, 1998). (The paper-and-pencil TOEFL is still administered in some international locations, and for institutional purposes in the USA.)

Computerized Language Listening Tests

Coniam (1996) observed that, with few exceptions, computerized testing of listening has received far less attention than has reading or structure. Coniam has pointed out that administering a traditional listening test imposes severe operational constraints, many of which can be addressed by computerized administration. In addition, a small scale experiment conducted by Coniam found a computerized listening test to be significantly correlated to a criterion listening measure, and to receive favorable comments from examinees.

It is notable that the computerized language tests discussed above are all used to make critical placement or proficiency decisions. All are large-scale tests that have received considerable funding. It would seem that far less has been done on a scale relevant to the everyday needs of the second and foreign language teacher. Given the availability of computers and commercial test-authoring software, classroom teachers should be able to take advantage of the benefits that computerized tests offer them and their students.

It appears that computerized listening tests have great potential for classroom practitioners; however, several challenges remain. Coniam considered recording audio files time consuming. Alderson (1990) critiqued computer test developers for not attempting any innovation in test method or content and for relying excessively on multiple-choice items. Issues of technology, measurement, language acquisition, and language teaching can be addressed by research in language testing.

PURPOSE

The purpose of the present research was to develop and compare computer-delivered and audiocassette/paper-and-pencil versions of a listening test. The test was a typical measure of

progress achievement of academic listening comprehension and vocabulary for high-intermediate level students of English as a Second Language (ESL). The test was based on Level III objectives for the English Listening/Speaking course. Objectives include: (1) sharpening listening skills through note-taking activities; (2) improving ability to understand and note main ideas and most supporting details of full-length presentations; and, (3) improving ability to make inferences and understand new ideas. Our underlying assumption, investigated here, is that the use of computer and audio technology for classroom progress tests can provide benefits of convenience, improved sound quality, and more, while providing measurement quality and validity at least as good as the paper forms of such tests.

METHODS

In the Spring of 1998, we developed a prototype class progress test in conjunction with several teachers in a university-affiliated ESL institute. The test was designed to assess the listening skills development of students in a Level III Listening and Speaking class, based on one specific unit, or chapter, from the class text.

After the initial item development, we constructed two parallel forms of the test, making both a computerized and a paper-and-pencil version of the exam. We pooled students across several classes and randomly assigned them to take either the computer- or the paper-version of the exam. After completing their exams, students in both conditions were also administered a short survey that asked a few demographic questions, along with questions about their attitude towards listening test administration mode. The two versions of the tests were then compared, to each other and to external measures of language listening ability, as a means of establishing the quality of the computerized progress test.

Sample

There were 28 students who took part in this study. There were both male and female students, from a wide variety of language/culture backgrounds. The students were enrolled in three separate classes (each taught by a different instructor), that were all Level III Listening and Speaking ESL classes. Level III students are expected to function at a fairly high level of ability, although they are not yet sufficiently proficient to participate in academic classes for native speakers. After successful completion of Level IV, many students would earn TOEFL scores that would satisfy university admissions requirements.

Description of the Prototype Progress Test

In order to compare the computerized listening test to a more familiar test format, both a paper-and-pencil and a computerized version of the unit exam were constructed. The item text was identical in the two versions, although a very few minor changes were necessary in the instructions. (For example, in the paper version, the instructions could read "Check the answer …", while in the computer version, the instructions would state "Click on the answer…").

The title of this text book unit was *Art, the Artist, and Society*. The skill of interest included simple vocabulary (related to art topics) and more advanced listening comprehension. The test was thus constructed in three sections; two short vocabulary sections and one listening comprehension section. (A copy of the paper version of the exam is provided in Appendix A.) One vocabulary section (Section 1) had audio prompts in which a definition was read aloud to the examinees; they then selected the word being defined from a set of vocabulary words. In the

other vocabulary section (Section 2) examinees saw a written vocabulary word and heard a sentence using the word in context; they then selected the correct definition for this word from a set of definitions. Finally, in the listening comprehension portion of the test, examinees heard a lengthy academic lecture on an art topic. They responded to two sets of questions about this lecture. The first set (Section 3 of the test) consisted of open-ended items that measured their ability to identify major themes of the lecture. The other set of listening comprehension questions (Section 4) used matching, and tested listening for details.

The audio portions of the two exams were recorded concurrently, to make their speed and clarity as similar as possible. That is, a speaker read the item scripts into microphones for both a cassette player and a computer at the same time. The cassette-based audio portions were designed to run sequentially, without stopping until the end of the test (timed pauses were recorded between each prompt). The computerized audio files were saved individually, and later imported into the computer test software. The computer version of the test was designed so that each examinee could play the sound file for a given item at his or her own initiative. In general, it is expected that computer audio files will be clearer than those stored and played using cassette tapes, due to the better technology for handling sound.

The computer-delivered exam was created using Authorware, version 3, to run on Windows-based machines, using a screen resolution of 640 x 480 pixels. The audio files were recorded at 11 Hz, 16 bit Mono. The total size of all audio files used in this exam was 3.98 MB (with the single long lecture comprising 1.26 MB of that total). The computers used to administer the exam were Pentium 200s, with 2.5 GB hard drives; they were also equipped with internal sound cards and headphones.

Description of the Survey

Students from both test conditions were administered a survey following their completion of the test (the survey questions are provided in Appendix B). The first section of the survey consisted of selected-response items. There were three selected-response items asked of examinees in both groups and one additional selected-response item asked only of the examinees in the computer group. For both groups, students were questioned as to (a) their level of computer experience, (b) their preference for taking either a computer version of a language test or a paper-and-pencil version with a cassette tape, and (c) which mode of test administration would make them more nervous. The students taking the computerized version of the test were also asked to compare the sound quality of the computer version with previously-administered paper-and-pencil/cassette tape versions of a language test. (This question did not apply to the students who took the paper-and-pencil version of the test because they had not been exposed to a computer administered test.)

In addition, students from both groups were requested to complete two open-ended questions, with students from the computer group completing a third open-ended question. Both groups were asked about (a) the general sound quality in their test mode and (b) their feelings about the ability (or inability) to control when the sound would start. The computer-group students were asked an additional question related to a computer navigation issue. Following the completion of the survey, several students from both testing conditions were randomly selected for follow-up oral interviews.

Description of Other Measures

Two external measures of language/listening ability were used in this study. These were the Comprehensive English Language Test (CELT) and the students' course grades for this semester's Listening and Speaking class (L/S grade).

CELT

The CELT is designed to measure intermediate to advanced levels of English proficiency of high school to adult aged persons who are non-native speakers of English. It is used as a placement tool by many English training institutes. Incoming students may be assigned to different course levels according to their scores. This instrument is also be used to assess progress: the CELT is sensitive enough to measure a learner's growth over the course of a semester. Using the CELT for pretest/posttest is relatively reliable and secure, since there are two forms of the CELT.

The test is composed of three sections: Listening, Vocabulary, and Structure. The Listening section consists of 50 items, delivered by audiocassette; it takes 40 minutes to complete. Students in the ELI program take this test on a semester basis, as a measure of their language learning progress. Their scores of the Listening component of the CELT were used as an external measure of language listening ability.

L/S grade

All of the students who participated in this study were enrolled in Level III Listening and Speaking (L/S) classes. The grade for the L/S classes were comprised of a number measures of both listening and speaking skills. These measures include tests, quizzes, in-class activities, and a final. This semester grade, while comprised of more than language listening, was also used as an external measure of language ability.

RESULTS

To investigate the validity of our computerized progress test of language listening skills, we compared the functioning of the computerized form to that of the paper form. Examinee performance on the two test forms (and the four sections making up each form) were compared directly to each other. A comparison of the performance of various examinee subgroups across the two test forms was also conducted. And, finally, a comparison of the extent to which each test form related to external measures of language listening ability was also conducted.

Sample Equivalency

The first issue addressed was the whether we could reasonably collapse across teachers and use the student as the unit of analysis. To accomplish this, we first examined students' English listening skills, as evidenced by the listening portion of the CELT test and the L/S grades. As noted in Table 1, a comparison of mean student CELT listening scores and L/S grades *between* teachers gives a strong indication that the students in the three classrooms were similar in average English listening ability.[1] As also depicted in Table 1, the comparison of the

[1] Because of the relatively low number of subjects—which is exacerbated by the number of subgroups—we were cautious of being overrelient on statistical testing. Instead, we tended to focus our attention on trends, to offer general evidence of the validity of the instrument.

mean student CELT scores and L/S grades—across the two modes of test administration for *each* teacher—gives a strong indication that the students taking the paper-and-pencil version of the language test and the students taking the computer version were also of equal ability.

Table 1

Mean Student CELT Scores and L/S Grades Scores By Teacher and Test Mode

	Teacher 1 Students			Teacher 2 Students			Teacher 3 Students		
Measures of English Skills by Test Mode	M	SD	n	M	SD	n	M	SD	n
CELT Score									
Computer	76.7	10.4	6	70.0	11.1	3	72.0	17.2	4
Paper	76.3	12.0	7	76.3	12.0	4	73.0	14.7	4
L/S Score									
Computer	86.0	8.3	6	86.7	2.5	3	82.8	11.5	4
Paper	88.0	6.7	7	81.3	7.6	4	85.8	3.3	4

Note: Maximum score = 100

Next, a comparison of mean student scores on the two modes of test administration between teaches was undertaken (see Table 2). While moderate differences are noted between teacher classrooms, these results were not unexpected (the test was based on a specific unit of instruction, and the three teachers were at different points in their coverage of the material when the test was given). What is important to note is that the differences, for the two modes of test administration, are relatively consistent between the classrooms (i.e., on *both* modes of test administration, students of teacher 1 scored better than the students of teacher 2 and students of teacher 2 scored better than students of teacher 3).

Based on the information presented, there does not appear to be a relationship between what went on in the different classrooms, the level of the students ability in the different classrooms, and the pattern of results across modes of test administration. Hence, it appears justifiable to use the student as the unit of analysis for the comparison of the two modes of test administration.

Table 2

Mean Test Scores By Teacher and Test Mode

	Teacher 1 Students			Teacher 2 Students			Teacher 3 Students		
Test Mode	M	SD	n	M	SD	n	M	SD	n
Computer	17.2	2.8	6	16.2	3.1	3	15.5	3.5	4
Paper	17.9	1.7	7	14.5	1.3	4	13.3	4.6	4

Note: Maximum score = 20

Comparability Across Administration Mode

The purpose of this project was to determine how a computerized language progress test, utilizing sound files, would perform in relationship to a paper-and-pencil test that utilized a cassette player. To accomplish this, the scores on the separate sections of the test and total test scores were compared between the two modes of administration.

While we were concerned about the over-reliance on statistical tests to analyze the data, it was nevertheless determined that the use of t-tests would add useful information in this instance. As Table 3 depicts, there were no statistically significant differences on any of the four sections or on the total score between the two modes of test administration (nor do the mean differences appear to be significant from a visual inspection). This provides evidence that the computer test mode is functioning similarly to the paper test mode.

Subgroup Analyses

Two subgroup analyses—for gender and language group—were conducted to investigate the possible presence of confounding effects. These effects may contribute to differential subgroup test performance resulting from factors unrelated to the specific skills and knowledge measured by the test.

First, language by mode of test administration was investigated. As noted in Table 4, there was relative consistency across test administration mode for the different language groups. This tends to support the analysis from the previous section, which showed an overall similarity between the two versions of the test. Second, gender by mode of test administration was investigated. Again, a relative consistency was noted across this grouping variable (see Table 5).

Table 3

Mean Section Scores and Total Test Scores By Test Mode

	Test Mode		
Test Section	Computer (n=13)	Paper (n=15)	t (p value)
Section 1			.49 (.64)
M	3.5	3.3	
SD	.78	.62	
Section 2			1.31 (.20)
M	3.3	2.8	
SD	.85	1.1	
Section 3			1.44 (.16)
M	6.8	5.9	
SD	1.4	1.7	
Section 4			1.31 (.21)
M	3.0	3.6	
SD	1.6	.90	
Test total			.54 (.71)
M	16.5	15.7	
SD	2.9	3.2	

Note: Maximum scores for Sections 1, 2 & 4 = 4; Section 3 = 8; Total test = 20

Table 4

Mean Test Scores Across Test Mode By Language Group

	Test Mode.	
Language	Computer	Paper
Spanish		
M	16.8	15.7
SD	2.6	2.7
n	4	7
Arabic		
M	13.3	15.4
SD	2.3	4.4
n	3	5
French		
M	19.0	18.0
SD	0	2.8
n	2	2
Japanese		
M	16.0	13.0
SD	4.2	--
n	2	1

Note: Two language subgroups—Italian and Korean—had only one member. Because they do not offer a cross-mode comparison, they were not included in the table.

Concurrent Validity

Finally, concurrent validity of the computer version of the progress test was addressed by examining the correlations between the two test administration modes and the available measures of English skills (CELT listening scores and L/S grades). A similar pattern of performance for the two progress test modes would suggest that the computer mode is comparable to the more familiar paper mode, in terms of validity. An examination of Table 6 confirms that the computer version relates to external measures of English skills at least at well as the paper version of the test.

Table 5

Mean Test Scores Across Test Mode By Gender

	Test Mode	
Gender	Computer	Paper
Male		
M	16.0	14.9
SD	3.2	3.6
n	8	10
Female		
M	17.2	17.4
SD	2.5	1.3
n	5	5

Table 6

Correlations Between Test Mode and Measures of English Skills

		Measures of English Skills	
Test Mode		CELT Score	L/S Grade
Computer	13	.74**	.21
Paper	15	.54*	.27

*p < .05. **p < .01.

Survey Analyses

Results of the survey were considered through descriptive and correlational analyses of the selected-response items and a qualitative analysis of the open-ended and interview questions. (All of the survey questions, for both groups, are provided in Appendix B.)

Selected-Response Items

Level of computer experience. The first selected-response item queried the students' level of computer experience. The options offered to the students were (a) beginner, (b) intermediate, and (c) advanced. As depicted in Figure 1, five students from the paper-and-pencil administration mode indicated a beginner level of experience, while six students indicated an intermediate level of experience and four indicated an advanced level. For the computer group, two students indicated that they were beginners, eight indicated intermediate status, and three indicated advanced status (level of computer experience was not significantly different between groups: $t = .53$, $p = .61$).

Correlations were examined, by group, between test scores and responses to this survey question. For the computer-version group, a significant effect was found between level of computer experience and test scores ($r = .70$, $p = .008$); this effect was not found to be significant for the paper-and-pencil group ($r = -.12$, $p = .68$).

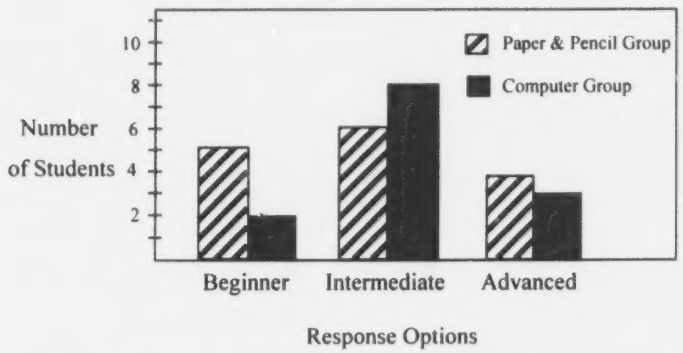

Figure 1. Responses to the the selected-response item *What is your level of computer experience?*

Preference of test-administration mode. The second selected-response item queried the students' preference for test-administration mode. For this question, the response options were (a) paper-and-pencil with a cassette tape, (b) it doesn't matter, and (c) computer. As noted in Figure 2, responses to this question were extremely similar for both testing groups. The majority of student responses were evenly split between *it doesn't matter* and the *computer version*, with only one respondent from each group indicating that they would rather take a pencil-and paper/cassette tape version of a language test. Additionally, for both groups, there was no statistically significant relationship between student test scores and their indicated test-administration preference (computer group: $r = .25$, $p = .41$; pencil-and paper group: $r = -.33$, $p = .27$).

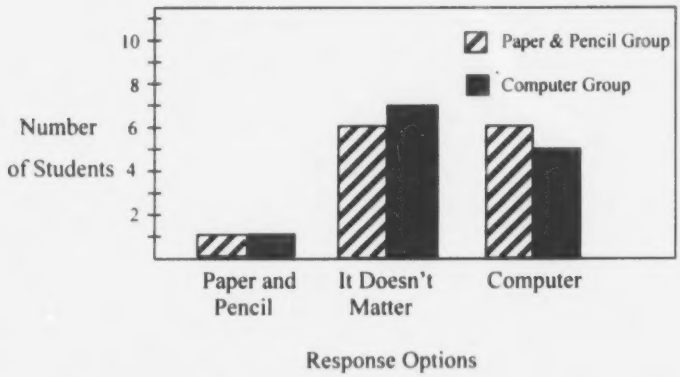

Figure 2. Responses to the the selected-response item *How do you prefer to take a listening test?*

Level of nervousness related to test-mode administration. The selected-response item *What kind of test makes you most nervous?* was asked to address students' comparative anxiety toward the two test modes. As with the previous selected-response item, the response options were (a) paper-and-pencil with a cassette tape, (b) it doesn't matter, and (c) computer. For the paper-and-pencil group, two students indicated that the paper-and-pencil mode of administration made them more nervous, two indicated that the computer mode made them more nervous, and

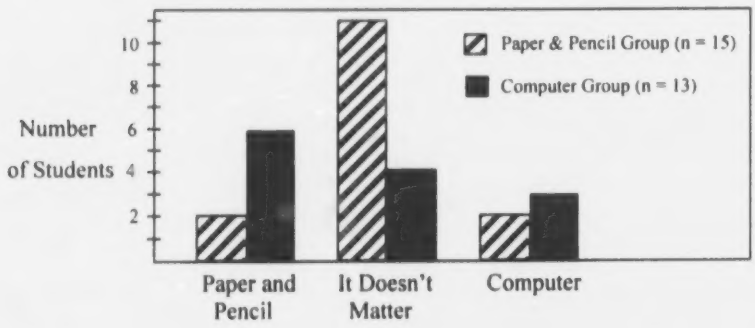

Figure 3. Responses to the the selected-response item *What kind of test makes you feel more nervous?*

eleven indicated that it didn't matter. For the computer group, six students indicated that the paper-and-pencil mode of administration made them more nervous, three indicated that the computer mode made them more nervous, and four indicated that it didn't matter (see Figure 3). As with the previous selected-response item, there was no statistically significant difference between the groups in their response to this question ($t = 1.26$, $p = .22$), nor was their a statistically significant relationship between students' test scores and their responses to this question (computer group: $r = .22$, $p = .46$; pencil-and paper group: $r = -37$, $p = .17$).

Comparison of sound quality between the two test modes. A question comparing the quality of sound between computer administration and paper-and-pencil/cassette administration was asked of the computer group only. In response to this question, a vast majority the students in this group—nine out of thirteen—indicated that the sound quality of the computer was *much* better than that of the cassette. Two others indicated that the sound quality of the computer was better than that of the cassette, one student indicated that there was no difference, and one student indicated that the cassette sound was better than the computer sound (see Figure 4). As with the two previous selected-response items, no statistically significant correlation was found between students' responses to this item and their test scores ($r = .11$, $p = .71$).

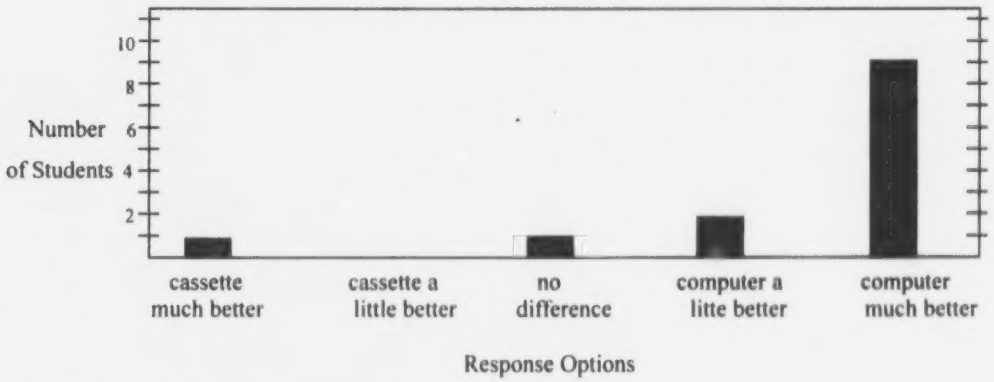

Figure 4. Responses to the the selected-response item *How clear is the sound in the two kinds of tests?* (computer group only)

Open-Ended and Interview Questions

Quality of sound. Both groups were asked, *Tell us how you feel about the test...how it sounded.* In the paper-and-pencil group, where the test was delivered via audiocassette, fifteen subjects answered this survey item. A total of four gave positive responses. While two of these were clearly about the cassette sound ("It sounds good, I can understand it."), the other two were ambiguous ("Interesting and it help [sic] me a lot."). One response was neither positive nor negative: "It depends on quality of cassette recorder." Most (ten subjects) of the paper-and-pencil group gave negative evaluations. While two of these comments were somewhat ambiguous (e.g., "nervous and uncomfortable"), the eight other negative responses were clearly

about the sound, such as "The sound was terrible" and simply "awful". It is interesting to note that this strong negative response was to an original recording, delivered in a small room, using a good cassette player. (The comments made by these learners of English have been quoted exactly, although spelling has been corrected.)

There were twelve participants in the computer group who responded to the same item. All twelve responses were positive and unambiguous. Typical of the comments was "It sounded very good. It has a good quality." While all computer group subjects stated that the audio quality was good, some identified other advantages to the test-taker. There were two subjects who remarked on concentration (e.g., "I think it permit [sic] me to concentrate much better."). Another two identified using headphones as beneficial (e.g., "It is good for me to be listened alone by headphone.") One participant noted "You feel you listen to the lecture alone." In follow-up oral interviews, several subjects expanded on their written comments. One remarked that it was good not to hear other people (because of using headphones); another said that she "hates the sound of papers turning" and the headphones kept her from being distracted.

Examinee control. The second survey question for the paper-and-pencil group was *Tell us how you feel about the test...that you can't control when to start the sound.* Out of the thirteen responses, one was positive ("I have control about myself.") and four were neutral (e.g., "It doesn't matter."). The other eight responses were negative. Six of these respondents made comments about their affective state, noting that they felt either nervous or stressed. One said, "I was felt all cooped up!"

For the computer group, the question read *Tell us how you feel about the computer test...that you controlled the sound.* Of the eight comments on this question, three were off-topic (such as "Is really clear the pronunciation, punctuation, and stressed word." and "It wasn't the first time."). The other five participants responded positively. One subject stated, "Is very good to control the sound, because you are ready when you feel that you are ready." Another said "I think this way is good for me. I always ready for next chapter [sic]."

Computer navigation. A third survey question was asked of the computer group only: *Tell us how you feel about the computer test...that you needed to work on two screens.* This question related to one section of the computer test, the lecture listening section, in which graphical limitations made it impossible to display all the related items on a single screen. Thus, the computer group had to alternate between two screens to answer the entire set of items related to the lecture. Additionally, software constraints meant that examinees could not respond to items while the lecture was playing. There were eight responses to this survey item. Three of the subjects' comments were ambiguous or off-topic (such as "I think the most important is to compare the pronunciation in real situation."), two were negative (e.g., "It's better if it's on one screen and we should be able to answer when we listen."), and three were positive (e.g., "No problem, all was easy.")

DISCUSSION

From a measurement standpoint, the computer version of the test appeared to perform as well as the paper-and pencil version. First, their were no statistically significant differences between the two groups, on either the total test scores or section scores. Second, there were no statistically significant effects, for either group, between test scores and (a) native language, (b) gender, (c) preference for mode of test administration, or (d) the level of nervousness towards test administration mode. Third, correlations between student test results and external measures

of English skills were comparable across the two test modes. These analyses provide evidence that the general validity of the computer test administration mode is quite comparable to that of the paper test administration mode.

A single statistically significant effect was noted. For the computer group only, it appears that a relationship may exist between level of computer experience and test score. This significant relationship appears primarily due to the small number of the subjects in the computer group who indicated a level of experience other than intermediate. As noted in the Results section, only three students from this group indicated advanced status, and only two indicated beginner status, with the majority of the students characterizing their computer experience as intermediate (eight out of thirteen). Further, the three individuals who indicated an advanced level of computer experience did quite well on the test (all scored 19 out of 20), while the two individuals who indicated their computer experience level as beginner did less well than average. Given this, it is difficult to determine if performance on the computer test was the result of examinees' language skills alone, or of their computer skills as well.

An untangling of these possible causes may be found in the fact that the relationship between computer experience and test results was not found to be statistically significant for the paper-and-pencil group. That is, for this group, the examinees with greater computer experience did not perform differentially better on their paper-and-pencil progress test. This tends to indicate that for this study, students' computer experience may have effected their performance on the computer test.

Further refinements in the computerized test may address this effect. For example, the instructions might be made clearer. However, writing for non-native speakers often results in overly long, stilted sentences because of the need to use basic English vocabulary. This is in conflict with principles of writing on-screen instructions, where clarity and brevity are of paramount importance. Balancing these two opposing guidelines is thus somewhat of a challenge. As a second option, a practice item could also be provided for each section of the test. Practice items are typically included in standardized tests, including computer versions of these exams. While these are not typically included in classroom tests, they might be reasonably added. In fact, however, these actions might not be necessary. It is anticipated that with only limited exposure to the computerized testing mode, students would become quickly familiar with the process, and the effect for computer experience would disappear.

Another issue encountered pertains to the different types of items that are typically available in most testing situations, and their associated constraints related to computerized testing. Open-ended item types, for example, may offer distinctive instructional and assessment merit, but are often difficult to score by the computer. Other item types, such as multiple choice, true-false, and cloze items, are easily scored by a computer, but often yield less information than might be provided by open-ended types of questions.

For certain course objectives, the use of open-ended item types is often necessary. In this present study, the course objectives—which included higher-order language skills—dictated the need for open-ended item types along with the more easily scored items. To address this, the progress test developed in this study included both multiple choice and open-ended questions, and the test report provided to the teacher for each student included the computer scored results of the multiple choice items, along with verbatim text of the student responses to the open-ended items.

Other issues in audio computer-based testing relate more directly to the technology used for computerized assessment. Coniam (1996) suggested that the time-consuming nature of recording computer audio files might be a problem. However, computers equipped with sound cards, recording software, and microphones have become readily available at reasonable prices (at least in America). For the present study, recordings were simultaneously made for computer and audiocassette. In both formats, recording did not prove to be a problem. Another technical issue concerns the size requirements for audio files. Audio files are much larger than text files, and thus require greater computer storage capacity. Given the much greater hard drive capacity currently available, this may not be a serious problem. The ability to master CD-ROMs is also becoming increasingly available. An audio-computerized test could be run from a CD-ROM rather than hard drive or diskette. While these and other technical issues need to be considered, they are likely to become less of a concern as hardware and software improvements continue to be made.

Finally, a most promising result of this study was the positive reaction by students who took the computer version of the test. Students in this group found the sound quality to be very good, and they appeared to appreciate being able to control when an audio prompt was played. These reactions are in marked contrast to those provided by students who took the paper mode of the exam. Those students found fault with the quality of the sound, and objected to the cassette-driven control of the prompts. Although it had seemed possible that students in the computer mode would object to the need to page between screens for items in the lecture section of the test, their comments were not markedly negative.

The overall results of our experiment were promising. Discussion with the teachers and students confirmed that the computer progress test offered distinct advantages. The teachers were positive about the potential for computerized administration and scoring of their class tests, while the students were enthusiastic about the improved sound quality in the computer mode and their ability to play the sound prompts at their own discretion. Our analyses, overall, found positive results for the computerized progress testing. While refinements and improvements to the computerized method of testing are necessary, we are encouraged by our progress to this point.

References

Alderson, C. (1990). Learner-Centered testing through computers: Institution issues in individual assessment. In H. de jong & D. Stevenson (Eds.), *Individualizing the assessment of language abilities*. Clevedon, England: Multilingual Matters.

Ariew, R., & P. Dunkel (1989). A prototype for a computer-based listening comprehension proficiency test. (Report, USDE International Research & Studies Program, Grant G008740406.) University Park, PA, 1989.

Bachman, L. (1991). *Fundamental considerations in language testing*. Oxford: Oxford University.

Coniam, D. (1996). Computerized dictation for assessing listening proficiency. *CALICO Journal* (13) 2 & 3, 73-85.

Educational Testing Service (1998). *Computer-Based TOEFL Score Users' Guide*. Princeton, NJ: Author.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London:

Madsen, H. (1991). Computer adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P. Dunkel (Ed.), *Computer assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Powers, D. (1986). Academic demands related to listening skills. *Language Testing, 3*(1), 1-38.

Richards, J. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 7*, 219-240.

Stevenson, J. & Gross, S. (1991). Use of a computerized adaptive testing model for ESOL/bilingual entry/exit decision making. In P. Dunkel (Ed.), *Computer assisted language learning and testing: Research issues and practice*. New York: Newbury House.

Additional Resources.

Alderson, C. (1998). Developments in language testing and assessment, with specific reference to information technology. *Forum for Modern Language Studies, 34*(2), 195-206.

Brown, J., D. (1997). Computers in language testing: Present research and some future directions. *Language Learning &Technology , 1*(1), 44-59 [on-line serial] Available on the World Wide Web: http://polyglot.cal.msu.edu/llt/vol1num1/brown/default.html [March 14, 1998].

Dunkel, P. (1991). Computerized testing of nonparticipatory L2 listening comprehension proficiency. An ESL prototype development effort. *Modern Language Journal, 75*, 64-73.

Flowerdew, J.(Eds), (1994). *Academic listening: Research perspectives.* Cambridge: Cambridge University Press.

Fulcher, G. *Resources in language testing page* [On-line]. Available on the World Wide Web: http://www.surrey.ac.uk/ELI/ltr.html.

Language Testing Research and Practice electronic discussion group. To subscribe, send e-mail to LISTSERV@LISTS.PSU.EDU. Leave the subject line blank; in the body, type the message subcribe LTEST-L <your firstname> <your lastname>.

Meunier, L. (1994). Computer adaptive language tests offer a great potential for functional testing. Yet, why don't they? *CALICO Journal, 4*(11), 23-39.

Appendix A

Paper-and-Pencil Listening Test
on *Art, the Artist and Society*

Instructions: Vocabulary 1

You will hear the definition of a word. You will read 4 vocabulary words. Check the word that matches the definition you heard. You can listen to the definition 2 times.

1. _____emotional
   _____outcast
   _____earthy
   _____primitive

2. _____exaggerate
   _____protest
   _____complex
   _____subjective

3. _____reasoned
   _____emerge
   _____prehistoric
   _____harmonious

4. _____economic
   _____controlled
   _____sophisticated
   _____emphasize

Instructions: Vocabulary 2

You will hear a sentence using a vocabulary word. You will read three definitions. Check the definition that matches the word in the sentence. You can listen to the sentence 2 times.

1. harmonious
   _____Pretty or attractive but not always necessary or useful.
   _____Recently developed styles that differ from traditional ones.
   _____When parts or pieces look good together and work well together.

2. sophisticated
   _____Having knowledge and understanding of complex subjects.
   _____Belonging to a society that has a simple way of life.
   _____Having styles that differ from traditional ones.

3. primitive
   _____Rough and simple, from an early stage of development.
   _____Having styles that differ from traditional ones.
   _____Pretty or attractive but not always necessary or useful.

4. decorative
   _____When parts or pieces look good together and work well together.
   _____When something looks attractive or ornamental.
   _____Recently developed styles that differ from traditional ones

Appendix A (Continued)

Instructions: Listening Comprehension
You will hear a lecture on Baroque and Impressionist art.
First, for each art style, write two things about the social context and the artistic context.
Second, you must match art styles with famous artists.

You will hear the lecture two times. Now, take a moment to read the questions.

Baroque - social/cultural context	Impressionism - social/cultural context
1. 2.	1. 2.
Baroque - artistic context	**Impressionism - artistic context**
1. 2.	1. 2.

Match the name of the artist with the correct art style.

Rubens	**Renoir**
1. Baroque 2. Impressionism	1. Baroque 2. Impressionism
Degas	**El Greco**
1. Baroque 2. Impressionism	1. Baroque 2. Impressionism

BEST COPY AVAILABLE

Appendix B

List of Survey Questions asked of Both Paper-and-Pencil and Computer Group Examinees

Selected-response questions

1. What is your level of computer experience?
    a. beginner
    b. intermediate
    c. advanced
2. Do you prefer to take a listening test…
    a. using paper-and-pencil with cassette tape
    b. it doesn't matter
    c. using a computer
3. What kind of test makes you feel more nervous?
    a. a paper-and-pencil test with a cassette tape
    b. it doesn't matter
    c. a computer test

computer group only:

4. How clear is the sound in the two kinds of tests?
    a. cassette tape sound was much better
    b. cassette tape sound was a little better
    c. there was no difference
    d. computer sound was a little better
    e. computer sound was much better

Open-ended questions

1. Tell us how you feel about the test…how it sounded.
2. *(computer group version)*
   Tell us how you feel about the test…that you controlled when to start the sound.
2. *(paper group version)*
   Tell us how you feel about the test…that you can't control when to start the sound.

computer group only:

3. Tell us how you feel about the test…that you needed to work on two screens.